Nucleic Acids Research, 2004, Vol. 32, Database issue ***D303–D306***
DOI: 10.1093/nar/gkh140

RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12

Heladia Salgado, Socorro Gama-Castro, Agustino Martínez-Antonio, Edgar Díaz-Peredo, Fabiola Sánchez-Solano, Martín Peralta-Gil, Delfino Garcia-Alonso, Verónica Jiménez-Jacinto, Alberto Santos-Zavaleta, César Bonavides-Martínez and Julio Collado-Vides*

Program of Computational Genomics, CIFN, UNAM. A.P. 565-A Cuernavaca, Morelos 62100, Mexico

Received September 12, 2003; Revised October 8, 2003; Accepted October 29, 2003

ABSTRACT

RegulonDB is the primary database of the major international maintained curation of original literature with experimental knowledge about the elements and interactions of the network of transcriptional regulation in *Escherichia coli* K-12. This includes mechanistic information about operon organization and their decomposition into transcription units (TUs), promoters and their σ type, binding sites of specific transcriptional regulators (TRs), their organization into 'regulatory phrases', active and inactive conformations of TRs, as well as terminators and ribosome binding sites. The database is complemented with clearly marked computational predictions of TUs, promoters and binding sites of TRs. The current version has been expanded to include information beyond specific mechanisms aimed at gathering different growth conditions and the associated induced and/or repressed genes. RegulonDB is now linked with Swiss-Prot, with microarray databases, and with a suite of programs to analyze and visualize microarray experiments. We provide a summary of the biological knowledge contained in RegulonDB and describe the major changes in the design of the database. RegulonDB can be accessed on the web at the URL: http://www.cifn.unam.mx/Computational_Biology/regulondb/.

INTRODUCTION

Escherichia coli has been a model organism since the beginning of molecular biology. Current post-genomic research in bioinformatics, network analyses and modeling, and system biology, can strongly benefit from studies in *E.coli*, given the large amount of accumulated knowledge of the molecular biology of this cell. It may be that this is the cell for which we know more about the function of its genes, its metabolism and transcriptional regulation. This knowledge is the foundation for the proposal within the International *E.coli* Alliance, to achieve in *E.coli*, as a long-term goal, the first whole-cell model (1). We contribute to this international effort with RegulonDB, the primary database of the major international maintained curation of original literature with experimental knowledge about the elements and interactions of the network of transcriptional regulation in *E.coli* K-12. It is a relational database containing mechanistic information about operon organization and their decomposition in transcription units (TUs), promoters and their σ type, binding sites of specific transcriptional regulators (TRs), their organization into 'regulatory phrases', active and inactive conformations of TRs, as well as terminators and ribosome binding sites. All this information is mapped onto the *E.coli* K12 chromosome. The database is updated constantly by searching in original publications, and is complemented by computational predictions. Every object has experimental evidence, and a direct link to the original publication via MedLine. Previous publications explain the initial relational design and subsequent modifications (2–5).

We estimate that we have ~20–25% of all predicted interactions of the network (see the summary of the increasing content of RegulonDB by year shown in Table 1). RegulonDB has been used in different types of analyses by the scientific community, such as predictions of regulatory sites (6) and operons (7–10); complementation of other databases, specifically, the mechanistic information gathered from the literature is included in EcoCyc (11); reconstruction of metabolic pathways with regulatory information (12); analyses of the connectivity and over-represented motifs in the regulatory network of *E.coli* (13–14); studies identifying objective criteria that characterize and define global regulators in *E.coli* (15); studies on the evolution of regulatory mechanisms (16–17), as well as analyses of microarray experiments (18).

The motivation to incorporate additional information comes from the fact that experimental research in *E.coli*, as in any

*To whom correspondence should be addressed. Tel: +527 77 313 2063; Fax +527 77 317 5581; Email: ecoli-t1@cifn.unam.mx

The authors wish it to be known that, in theirr opinion, the first five auhtors should be regarded as joint First Authors

Table 1. Outline of RegulonDB information gathered by year

Object	1997	1998	1999	2000	2001	2002	2003
Regulons	99	83	83	165	166	172	179
Regulatory interactions	533	433	433	642	935	990	1119
Sites			406	469	750	812	950
Products			4405	4405	4405	4405	4408
RNAs			115	115	115	115	116
Polypeptides			4207	4290	4290	4290	4292
Transcriptional regulators[a]	99	83	83	165	166	170	179[b]
Genes	542	456	4405	4405	4405	4405	4408
Transcription units	292	230	374	528	657	694	747
Promoters	300	239	432	624	746	783	860
Effectors	35	36	36	36	66	66	67
External references	2050	2011	4394	4704	4943	5053	5224
Synonyms		681	3525	3525	3525	3544	3578
Terminators			40	86	106	108	118
RBSs			59	98	133	134	153
Conformation of transcriptional regulators				83	201	203	221

[a]The term 'Protein complex' has been change to 'Transcriptional regulator'.
[b]Total of 318 transcriptional DNA binding regulators, of which 179 have experimental evidence and the rest have been predicted based on their helix–turn–helix DNA binding motif (16).

other organism, goes well beyond knowledge about the molecular biology involved in regulation and transcription. Physiological and genetic studies add a rich layer of knowledge about the internal structure of the cell. There is, for instance, a large number of publications describing the effect in the expression of specific genes when changing the growth conditions of the cell, specifically experiments studying the effects of deletions of regulatory genes. This genetic and physiological information provides knowledge without necessarily specifying the corresponding molecular mechanisms.

Having this information expands the utility of RegulonDB. For instance, it can be used to compare and validate microarray experiments (18). Computational genomics has grown in methods and goals, moving from a sequence-centered approach to one where regulatory networks and interactions have become the main focus. Understanding the regulatory network will be crucial in the future goal of modeling, *in silico*, the behavior of *E.coli* as an entire cell (1).

In the following, we describe how growth conditions are modeled in the databases and then summarize the computational changes and additions to the database.

RESULTS

Gene expression changes as a function of growth conditions in RegulonDB

Free-living bacteria have to maintain a constant monitoring of extracellular physicochemical conditions in order to respond and modify their gene expression patterns accordingly. A series of genes whose products are involved in sensing and incorporating the different nutritional elements, as well as products sensing the concentration of toxic elements, are present in *E.coli*. These sensing systems are connected through metabolic intermediates to the transcriptional machinery, which in turn modifies the expression of genes whose products are involved in the response and adaptation to the corresponding changes in the environment. For the past 2 years, we have been collecting and organizing, from the original literature, information about different growth conditions and the associated observed effects in the transcription of *E.coli* genes. Since the first published version of RegulonDB (2), we described in the relational design the modeling of physiological conditions and their connection to the transcriptional machinery. However, as mentioned in that paper, we were not then involved in gathering such types of information.

After an analysis of several different conditions and systems involved, we decided to implement a model where the following properties and descriptions are considered essential: (i) a general or global condition; (ii) the control condition; (iii) the specific experimental condition; (iv) the growth media used; (v) the genes affected; and (vi) the effect of the experimental condition in the expression of the affected genes (induced, repressed or no effect). Since every added object in RegulonDB is supported by associated evidence and literature citation, we had to implement a set of criteria to classify the evidence concerning different levels of expression of genes.

To quantify gene expression, by far the most frequent methodology is that of transcriptional fusion. These studies provide quantitative information easy to classify. We incorporate as affected genes those with an expression change of least a 2-fold increase or decrease. Otherwise, genes are added to the database considered cautiously as genes with 'no effect' or no change in expression under the specified condition. In a small fraction of cases there is no quantitative information on the level of expression of the affected gene and, therefore, its classification is not straightforward. In those cases the curator's criterion is essential. The classification of the level of expression depends on the authors' statements, the visual inspection of the spots in the figures in the publication, as well as, ideally, additional evidence in other publications.

Whenever available, additional information is incorporated, i.e. mechanistic properties that are already part of RegulonDB: (i) the transcription unit to which the gene belongs, the associated promoter and terminator; (ii) the regulatory protein that is involved; (iii) the set of sites in the DNA involved in regulation of transcription; (iv) the allosteric conformations and associated effectors involved; as well as (v) the intermediate metabolites or proteins that participate in the

Table 2. Summary of environmental conditions gathered in RegulonDB 4.0

Object	Total
Global conditions	16
Specific conditions	83
Genes	327
Transcriptional regulators	32
Conformations of transcriptional regulators	40
Promoters	116
Transcription units	57
Intermediates	30
Evidences, methods	2, 5
References	228

regulatory sensing mechanism. The design and discussion of the potential applications of this corpus of knowledge is presented in more detail in a separate paper (19).

Table 2 summarizes the information that we have gathered up to September 2, 2003 concerning physiological conditions and their effect on the transcription of genes. The numbers in this table account for unique cases, thus 327 genes have information about their expression in 83 different conditions. Since there is information for genes affected in different conditions, these genes are described a total of 679 times with their associated specific conditions.

Computational changes to the interface

We have changed the web interface so that the main menu remains fixed throughout navigation. For instance, the ZoomTool that displays the whole genome is now shown without invoking an additional external window. We have added a new selection by functional class within the graphic display. A very useful navigation feature in the analyses of transcriptome data is the new capability of taking a file with a list of genes and getting their display in the circular genome. GETools, a suite of programs linked to the database, was specifically designed to analyze, generate graphic displays and extract information from RegulonDB, from an input based on microarray files (20).

Alignments and matrices for each transcriptional regulator have been updated, and their automatic update as new sites from the literature accumulation has been implemented. The process begins by getting all the regulatory binding sites with experimental evidence, then, the program CONSENSUS 5c (21) is applied to generate the corresponding weight matrix. We get the first matrix of the second cycle, where all the sequences are included. This matrix and the program PATSER 3b (22) are used to score these same known sites. From this scoring, we define alternative thresholds available for the user, to search for similar sites in other DNA sequences. RegulonDB users can obtain these data by querying for 'Transcriptional Regulator'.

There are two ways the user can access the information on growth conditions that affect specific genes, either through a list of conditions available in the main page, or by searching for individual genes. Furthermore, we have added links to the OU microarray database (http://www.ou.edu/microarray/Macroarray/). RegulonDB is also now linked to Swiss-Prot and Swiss-Prot has links to RegulonDB.

DISCUSSION

The information on the effect of growth conditions on gene expression will be of great value in defining and modeling functional modules in cellular physiology. Metabolic intermediates and environmental signals, functioning as allosteric effectors of transcriptional factors, are additionally available in RegulonDB. Together, this information will enable a more complete description of sets, or modules, of genes as they are expressed in *E.coli* in response to different environmental conditions.

An example of the use of the knowledge gathered in the database is the comparison of what RegulonDB would predict in terms of expression profiles, and what is observed in microarray experiments (19).

We have also made a proposal of diagnostic criteria to identify global regulators, where we have shown that global regulators are active in a larger number of different growth conditions than specific or dedicated regulators. This observation enriches the original requirement of global regulators to regulate genes that belong to different metabolic pathways (15).

The current expansion of data gathered and organized in RegulonDB will reinforce and contribute to the efforts of the international community in the long-term goal of modeling of the full *E.coli* cell (1).

We kindly ask users of RegulonDB to cite this article.

ACKNOWLEDGEMENTS

We acknowledge Rosa María Gutiérrez-Ríos and Mónica Peñaloza-Spínola their participation in discussions on growth conditions, and Víctor del Moral and Romualdo Zayas for their computer support. This work was supported by NIH grants GM62205-02 and 1-R01-RR07861.

REFERENCES

1. Holden,C. (2002) Alliance launched to model *E. coli*. *Science*, **297**, 1459–1460.
2. Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) RegulonDB: a database on transcription regulation in *Escherichia coli*. *Nucleic Acids Res.*, **26**, 55–60.
3. Salgado,H., Santos,A., Garza-Ramos,U., van Helden,J., Díaz,E. and Collado-Vides,J. (1999) RegulonDB (version 2.0): a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res.*, **27**, 59–60.
4. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate,D., Blattner,F.R. and Collado-Vides,J. (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in *Escherichia coli*. *Nucleic Acids Res.*, **28**, 65–67.
5. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate,D., Díaz-Peredo,E., Sánchez-Solano,F., Pérez-Rueda,E., Bonavides-Martínez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): Transcriptional regulation and operon organization in *Escherichia coli K-12*. *Nucleic Acids Res.*, **29**, 72–74.
6. Tan,K., Moreno-Hagelsieb,G., Collado-Vides,J. and Stormo,G.D. (2001) A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, **11**, 566–584.
7. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.
8. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
9. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) Operon conservation from the point of view of *Escherichia coli* and inference of functional

interdependence of gene products from genome context. *In Silico Biol.*, **2**, 87–95.
10. Zheng,Y., Szustakowski,J.D., Fortnow,L., Roberts,R.J. and Kasif,S. (2002) Computational identification of operons in microbial genomes. *Genome Res.*, **12**, 1221–1230.
11. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.
12. Covert,M.W., Schilling,C.H. and Palsson,B. (2001) Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.*, **213**, 73–88.
13. Oosawa,C. and Savageau,M.A. (2002) Effects of alternative connectivity on behavior of randomly constructed Boolean networks. *Physica D.*, **170**, 143–161.
14. Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.*, **31**, 64–68.
15. Martínez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.*, **6**, 482–489.
16. Pérez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli*. *Nucleic Acids Res.*, **28**, 1838–1847.
17. Babu,M.M. and Teichmann,S.A. (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1234–1244.
18. Gutiérrez-Ríos,R.M., Rosenblueth,D.A., Loza,J.A., Huerta,A., Glasner,J.D., Blattner,F. and Collado-Vides,J. (2003) Regulatory network of *Escherichia coli*: Consistency between literature knowledge and microarray profiles. *Genome Res.*, **13**, 2435–2443.
19. Martínez-Antonio,A., Salgado,H., Gama-Castro,S., Gutiérrez-Ríos,R.M., Jiménez-Jacinto,V. and Collado-Vides,J. (2003) Environmental conditions and transcriptional regulation in *Escherichia coli*: A physiological integrative approach. *Biotechnol. Bioeng.*, **84**, 743–749.
20. Huerta,A.M., Glasner,J.D., Jin,H., Blattner,F.D., Gutiérrez-Ríos,R.M. and Collado-Vides,J. (2002) GETools: Gene Expression Tool for analysis of transcriptome experiments in *Escherichia coli*. *Trends Genet.*, **18**, 217–218.
21. Stormo,G.D. and Hartzell,G.W.,3rd (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
22. Hertz,G.Z., Hartzell,G.W.,3rd and Stormo,G.D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.